Evaluative Conditioning Effects Are Modulated By The Nature of Contextual Pairings

Sean Hughes, Yang Ye, & Jan De Houwer

Ghent University

Authors Note

SH, YY, and JDH, Department of Experimental Clinical and Health Psychology, Ghent University. This research was conducted with the support of Grant BOF16/MET_V/002 to JDH. Correspondence concerning this article should be sent to sean.hughes@ugent.be.

Abstract

Across two studies participants completed a learning phase comprised of two types of trials: context pairing trials in which two (valenced or non-valenced) words were identical or opposite to one another and evaluative conditioning (EC) trials in which a CS was paired with a US. Based on the idea that EC occurs because CS-US pairings function as a symbolic cue about the relation between the CS and the US, we hypothesized that the nature of context pairings (identical or opposite) might moderate EC effects. Results indicate that identity-based context pairs led to typical assimilative explicit and implicit effects whereas opposition-based pairs led to attenuated effects. Implications and different accounts of our findings are discussed.

Keywords: Evaluative Conditioning, Symbolic, Relational Qualifier, Implicit Evaluation

Evaluative Conditioning is Modulated by the Nature of Contextual Pairings

Evaluation is at the core of human psychology. It not only guides our judgments and decisions but often dictates how we treat our friends and family, as well as novel individuals, social, religious, and ethnic groups. Evaluations are assumed to bias what we remember, influence the politicians we vote for, musicians we listen to, and the products we consume. It is therefore crucial that we understand how, when, and why evaluations are established and what factors play a role in their change.

Many researchers have focused on Evaluative Conditioning (EC) - a change in liking due to the pairing of stimuli - as a means of establishing and manipulating evaluations. In a typical EC study a neutral conditioned stimulus (CS) acquires the valence of a positive or negative unconditioned stimulus (US) with which it was previously paired. For example, contiguous presentations of an unknown brand product with pleasant images can result in that product being evaluated positively whereas pairing it with negative images results in it being evaluated negatively (see Hofmann, De Houwer, & Perugini, Baeyens, & Crombez, 2010).

Although EC has traditionally been viewed as a rather "primitive" or simple form of learning (e.g., Briñol, Petty, and McCaslin, 2009) we recently introduced a new 'symbolic' perspective on this phenomenon (De Houwer & Hughes, 2016). This perspective argues that, early on in their development, humans gain access to a symbolic learning pathway, one that enables them to relate stimuli in a diverse number of ways. Once this ability is acquired, they can come to treat virtually any proximal event in the environment as a symbol (or cue) which can - in turn - influence their behavior. These cues can take many forms: from words and sentences, to musical or mathematical notation, physical objects (e.g., a red traffic light), or even gestures such as a wink of the eye or nod of the head. Put another way, stimuli in the

environment can function as symbols on the basis of which symbolic meaning is constructed.[1]

If humans are capable of imbuing stimuli with symbolic meaning (e.g., acting-as-if the word 'snake' stands for an actual snake), then it seems reasonable to assume that they can also imbue environmental *regularities* with such meaning as well. An environmental regularity refers to "*all states in the environment of the organism that entail more than the presence of a single stimulus or behavior at a single moment in time*." (De Houwer, Barnes-Holmes, & Moors, 2013, p. 634). For instance, the environment can be arranged to create regularities such as the repeated presentation of a single stimulus (e.g., mere exposure) or relationships between stimuli and actions (e.g., approach-avoidance training). We propose that stimulus pairings - the regularity at the core of EC - represents yet another that can convey symbolic meaning. Our hypothesis is that the pairing of stimuli changes liking because humans respond to those pairings as a contextual cue symbolizing that the CS and US are related in a certain manner. From this perspective, EC research provides unique information about the way that symbolic meaning construction on the basis of stimulus pairings gives rise to changes in liking (De Houwer & Hughes, 2016).

EC as a Symbolic Phenomenon

Conceptualising EC as a symbolic phenomenon has both heuristic value (it can accommodate a number of recent findings in the literature) and predictive value (it leads to several new empirical possibilities). First, if pairings do function as a contextual cue signalling how the CS and US should be related, and people have a learning history of doing

[1] Note that our definition of a 'symbol' and 'symbolic meaning construction' is situated at a very abstract level of analysis. We view *symbols* as stimuli which are in some ways functionally substitutable for other stimuli in the environment (i.e., something that can stand for something else) and *symbolic meaning construction* as the transformation of a proximal event into a symbol by organisms with the aforementioned ability. By operating at a high level of abstraction, we hope to obtain new insights and reach consensus that might not be achievable when operating at the level of specific theories about learning and liking. That said, there are theories at both the functional (e.g., Hayes, Barnes-Holmes, Roche, 2001) and mental levels (e.g., Deacon, 1997) which offer insight into the origins and nature of symbols and symbolic meaning construction and which seem compatible with the general position we forward here.

so, then there may be some 'default' symbolic meaning that they attribute to stimulus pairings. Based on the EC literature (e.g., Hofmann et al. 2010) and findings elsewhere in learning psychology (e.g., Hughes, De Houwer, & Barnes-Holmes, 2016), we hypothesize that the 'default' symbolic meaning of pairings is 'similarity' – namely – that the CS and US are similar along a particular dimension (e.g., valence). Similarity relations typical lead to the assimilative effects seen in the EC literature wherein a CS acquires the same valence as a US. The key point here is that, for organisms with access to the symbolic learning pathway, stimulus pairings may function in much the same way as the expression "*is similar to*" in the instruction "*A is similar to B*" (i.e., they symbolize that two stimuli share certain properties; De Houwer & Hughes, 2016).

Second, if humans do treat pairings as a symbolic cue upon which meaning is constructed, then it should be possible to manipulate the meaning that pairings convey. Just as the symbolic meaning of individual stimuli can vary across contexts (e.g., the letter-string 'bad' means 'bathtub' in Dutch and 'negative' in English), it might also be that the symbolic meaning of stimulus pairings also varies over contexts. In other words, the symbolic meaning of pairings is not entirely fixed: in one context, pairings could symbolize that the CS and US are similar to one another and yet in another context, pairings could symbolize that the CS and US are opposite or that they are causally or hierarchically related to one another. This idea is consistent with recent EC work showing that the impact of pairings on liking can be modified in a variety of ways. Several researchers have introduced relational qualifiers such as the words 'enemy' or 'loathes' during CS-US pairings and found that this can lead to contrastive EC effects (i.e., the CS acquires the opposite valence of the US it was paired with; Fiedler & Unkelbach, 2011; Förderer & Unkelbach, 2012). Others have provided explicit instructions before or after the EC phase stating that paired words have a particular meaning (e.g., synonyms or antonyms) and found assimilative and contrastive effects, respectively

(Hu, Gawronski, & Balas, 2017; Zanon, De Houwer, Gast, & Smith, 2014). Still others have asked participants to complete judgement tasks that focused their attention on the relative value of the CS or US and produced similar outcomes (Unkelbach & Fiedler, 2016) or primed a certain relation (similarity or difference) prior to the EC phase (Corneille, Yzerbyt, Pleyers, & Mussweiler, 2009). For instance, Corneille et al. initially primed participants to focus on either the perceptual similarities or differences between stimuli and then administered a (purportedly unrelated) learning phase in which CSs were paired with USs. They found that participants who were primed with the goal of processing perceptual similarities between stimuli showed larger EC effects than those primed with the goal of processing their perceptual differences. Finally, a combination of instructions about, and then exposure to, paired events can lead to contrastive EC effects as well (e.g., Moran & Bar-Anan, 2013; Peters & Gawronski, 2011).

For the main part, the aforementioned work interpreted these contrastive EC effects as being due to the impact of a symbolic message that is delivered *on top of* pairings and often assumed that implicit evaluations would only be influenced by pairings whereas the symbolic message would exert an exclusive influence on explicit evaluations. Hence, the symbolic message and the pairings were seen as distinct sources of changes in evaluations, with pairings being inherently non-symbolic in nature. We take a different perspective. Drawing on the symbolic account, we hypothesize that the aforementioned effects can arise because relational qualifiers, instructions, contrastive judgements, and priming tasks transform the symbolic meaning of the pairings themselves. If this is the case, and pairings do have a default symbolic meaning, then it should be possible change this meaning in multiple ways, not only by providing verbal information about the pairings but also by adding contextual

pairings[2]. More specifically, presenting pairs of stimuli that are opposite to one another could signal that, within that broader context, pairings is a cue indicating that stimuli are opposite to one another (along some dimension). Hence, if a neutral CS is paired with a valenced US in that same context then this could result in an attenuated or even reversed EC effect. Likewise, presenting contextual pairs of identical stimuli might strengthen the default symbolic meaning of pairings and thus lead to stronger EC effects. In short, a symbolic conceptualisation of EC leads to the novel prediction that EC depends on the relational properties of stimulus pairs that occur within the same context as the CS-US pairs. One way to test this hypothesis is to examine if contextual pairings moderate EC effects.[3]

The Current Research

Towards this end, we set out to moderate the direction and magnitude of EC effects by means of context pairings. During a learning phase participants were exposed to two different types of trials: EC trials in which a neutral nonsense word (CS) was paired with valenced words (US) and context trials in which two words were presented that were either identical (e.g., *day-day, up-up, cold-cold*) or opposite in their meaning (e.g., *day-night, up-down, cold-hot*). We assessed liking via self-report ratings and an Implicit Association Test (IAT). We added an IAT as it is assumed to reflect more automatic instances of evaluation that can influence behavior in unique ways (e.g., De Houwer, Teige-Mocigemba, Spruyt, &

[2] The priming manipulation of Corneille et al. (2009) could been seen as another way of altering the symbolic meaning of pairings via contextual manipulations (i.e., they directed attention towards the similarities or differences between one set of paired events [pictures] and then introduced a second set of such events [CS-US pairings]). Although the priming manipulation could have exerted an impact because of this reason, it is important to note that this manipulation was designed explicitly with the aim of changing the perceptual processing of stimuli. Hence, the effects of the priming manipulation could also have merely influenced perceptual processes.

[3] At the mental level of analysis our symbolic view is distinct from, but certainly compatible with, propositional models of EC. Given that our symbolic view considers pairings as cues which specify how stimuli are related, and that propositional representations are necessary in order to encode [higher-order] relational information in a mental system, it seems likely that propositions are a necessary component for symbolic meaning construction to occur. Critically, however, our symbolic view extends beyond propositional models by highlighting that pairings themselves can function as symbols (not just the CSs and US which are paired) and that pairings can also influence the relational content of the propositional belief.

Moors, 2009). If EC effects can be moderated by changes in context, then we would expect to observe larger EC effects when context trials involve identical compared to opposite stimuli.

Finally, past work on oppositional relational qualifiers and instructions has tended to reveal attenuated rather than fully reversed EC effects. This outcome has driven theorising in the area and led to a number of competing accounts (see Moran, Bar-Anan, & Nosek, 2016a). The symbolic perspective outlined above speaks to this issue as well. As we outline in the General Discussion, there are two versions of the symbolic account (one strong and one weak) that differ in the type and number of assumptions that they make. The *weak symbolic perspective* argues that - under certain conditions - pairings can function as a symbol and thus influence the magnitude and evaluative direction of EC effects. Yet in other cases pairings function as a mere *proximal cause* of the change in liking (i.e., the impact of the pairings on liking merely depends on their spatio-temporal [and not their symbolic] properties). In contrast, the *strong symbolic perspective* argues that pairings always function as a symbol: once an individual has acquired the ability to respond symbolically, stimuli in the environment are always imbued with such meaning (i.e., they never function as mere proximal causes). Therefore, according to the weak symbolic perspective, attenuation effects emerge when pairings function as a symbolic cue for some people and as a mere proximal cause for others. From the weak symbolic perspective, it is also possible that pairings function as a symbolic cue for everyone but that the meaning of that cue either differ *within-subjects* (the same person can simultaneously treat some pairings as a cue for similarity and other pairings as a cue for opposition) or *between-subjects* (one person can treat it as a cue for similarity whereas others can treat it as a cue for opposition). From a strong symbolic perspective, attenuation effects emerge when pairings have multiple meanings for the same person (within-subjects explanation) or a single meaning which differs between people (between-subjects explanation). Both accounts would predict that fully reversed effects

emerge when pairings function as a symbolic cue, and when the meaning of that cue is consistent within and between participants (i.e., pairings is a cue for opposition).[4]

Experiment 1

Method

Participants and Design. 100 participants (49 women) ranging from 18 to 49 years (*M* = 30.2, *SD* = 7.7) participated online via the Prolific Academic website (https://prolific.ac) in exchange for a monetary reward (€1.25). The experiment was programmed in Inquisit 4.0 and hosted via Inquisit Web (Millisecond Software, Seattle, WA). The experiment involved a single factor between-subjects design (*Context Pair Type*: same vs. opposite) with self-reported ratings and IAT scores as the main dependent variables. Three additional method variables were manipulated between participants: evaluative task order (self-reports vs. IAT first), IAT block order (learning [EC] phase consistent vs. inconsistent first), and stimulus assignment (CS1 with positive/negative USs). The sample size was determined prior to data collection in a convenience sampling manner. Note that the study designs, data-analysis plans, scripts, and data for both experiments are available on the Open Science Framework website (https://osf.io/bn2q3/). We report all manipulations and measures used in the study.

Materials. Two nonsense words served as CSs (*Ambik* and *Safrom*). Ten positive (*healthy, joy, freedom, friendship, honesty, beautiful, pretty, love, happiness, delightful*) and ten negative adjectives (*miscarriage, torture, cancer, disgusting, tumor, pain, rape, agony, hate, sickness*) served as USs. The words up, down, night, day, rich, poor, summer, winter,

[4] We ran two initial pilot studies (*n* = 41 and *n* = 60) that attempted to influence EC effects by manipulating the nature of context pairs (words vs. pictures), and context pair training (blocked vs. interspersed). Results indicated that implicit and explicit evaluations were consistently biased by context pairs, such that assimilative EC effects emerged when the context pairs signaled that pairings were a cue for sameness whereas attenuated effects emerged when pairings were a cue for opposition. Experimental scripts, raw data, and analyses for these two pilot studies can also be found on the OSF website (https://osf.io/bn2q3/).

old, young, sick, healthy, strong, weak, fast, slow, right, left, hot, cold, black, white, dead, alive, before, after, even, and odd were presented during context pair trials.

During the IAT the two CSs served as labels for the target stimuli and the words "Good" and "Bad" served as labels for the attribute stimuli. Eight positively valenced and eight negatively valenced adjectives served as attribute stimuli (*fantastic, great, magnificent, lovely, excellent, wonderful, amazing, super* versus *terrible, nasty, poor, horrible, hideous, awful, rotten, unpleasant*) and the two CSs in different fonts and orientations served as the target stimuli.

Procedure

There were three experimental phases: a learning phase, evaluative measures, and exploratory questions.

Learning phase. Prior to the learning phase participants were informed that they would either see two words or images and that they should pay attention to these at all times. Overall they encountered eight blocks that each contained 10 trials (80 trials total). The task was comprised of two different types of blocks (i.e., four blocks containing context trials and four blocks containing EC trials). Context trials involved the simultaneous presentation of two words for 3000ms. Participants in the *identical context pairs* condition encountered identical words (e.g., *Old-Old, Up-Up*) whereas their counterparts in the *opposite context pairs* condition were exposed to words that were opposite in meaning (e.g., *Old-Young, Up-Down*). Thereafter both stimuli disappeared, and following a 1000ms inter-trial interval (ITI), the next context pair was presented. EC trials involved the presentation of CS1 along with one of ten different positively valenced adjectives or CS2 along with one of ten negative adjectives for 3000ms. Both context and CS-US pairs were presented in random order within their respective blocks. The onscreen location of stimuli was always the same within a given block. That said, we (a) varied the location of the stimuli across blocks (e.g., the contextual

stimuli or CSs and USs were either presented parallel, vertically, or diagonally adjacent to one another) and (b) used the same stimulus location parameter in consecutive context and EC blocks (i.e., if the context stimuli were presented parallel to one another then so too were the CSs and USs in the following block). We adopted such a design in the hope that it would maximize generalization from context trials to EC trials by highlighting that stimuli in both types of trials were presented in the same temporal and spatial manner. Finally, we presented the context and EC blocks in sequential order such that a context block was always followed by a EC block of trials.

Self-reported ratings. Self-reported ratings were assessed using four different semantic differential scales. On each trial, one of the two CSs was presented and participants were asked to indicate their general impression of the stimulus using a scale ranging from −10 to +10 with 0 as a neutral point. The four end-points of the scales were as follows: *Negative-Positive, Pleasant-Unpleasant, Good-Bad, I Like It-I Don't Like It*. A mean evaluative rating was calculated for each CS by averaging scores from these four scales.

IAT. Automatic evaluative responding was assessed using an IAT. Participants were informed that a series of words would appear one-by-one in the middle of the screen and that their task was to categorize those items with their respective target (CS1 or CS2) or attribute categories ('Good' and 'Bad') as quickly and accurately as possible. They were told that the two items they had previously encountered (targets) as well as the words "Good" and "Bad" (attributes) would appear on the upper left and right sides of the screen and that stimuli could be assigned to these categories using either the left ('E') or right keys ('I'). Each trial started with the presentation of a target or attribute stimulus in the middle of the screen. If the participant categorized the word correctly – by selecting the appropriate key for that block of trials – the stimulus disappeared from the screen and the next trial began. In contrast, an incorrect response resulted in the presentation of a red "X" which remained on-screen until

the correct key was pressed. Overall, each participant completed seven blocks of trials. The first block of 20 practice trials required them to sort CS1 and CS2 into their respective categories, with CS1 assigned to the left ('E') key and CS2 to the right ('I') key. On the second block of 20 practice trials, participants assigned positive words to the "Good" category using the left key and negative words to the "Bad" category using the right key. Blocks 3 and 4 (20 and 40 trials, respectively) involved a combined assignment of target and attribute stimuli to their respective categories. Specifically, participants categorized CS1 and positive words using the left key and CS2 and negative words using the right key. The fifth block of 20 trials reversed the key assignments, with CS2 now assigned to the left key and CS1 to the right key. Finally, the sixth and seventh blocks (20 and 40 trials respectively) required participants to categorize CS1 with negative words and CS2 with positive words.

Exploratory questions. We assessed whether participants were hypothesis aware using the following questions: *"What do you think the aim of this experiment was?*" and "*How do you think the experiment achieved this*?". We also assessed for context pair awareness ("*During the experiment, did you notice that we sometimes presented words that were SIMILAR/OPPOSITE to one another together onscreen?*"), and context pair influence ("*Did this influence how you responded to CS1 and CS2*"). A manipulation check was also included to check whether participants physically recorded the contingencies operating during the learning phase. We also assessed for CS-US contingency memory, how confident participants were in their evaluations, and asked them to complete a behavioral choice task. Many of these variables were registered purely for exploratory purposes and will not be discussed further in the paper.

Results

Data Preparation

In-line with IAT data treatment elsewhere in the literature (e.g., Smith, De Houwer, & Nosek, 2013), data-exclusion involved removing participants who had IAT error rates above 30% across the entire task or above 40% for any one of the four critical blocks (n = 5), or who responded faster than 400ms on more than 10% of trials (n = 3). This led to a final sample of 92 participants.[5]

Analytic Strategy

To determine whether self-reported ratings and automatic evaluative responses towards CS1 and CS2 (*dependent variables*) differed as a function of the type of context pairs presented (same vs. opposite) (*independent variable*), a series of ANOVAs and post-hoc t-tests were carried out on the rating and IAT data.

Preliminary Analyses

Descriptive analyses for hypothesis and context pair awareness/influence as well as contingency memory can be found in Table 1.

Hypothesis Testing

Self-reported ratings. Mean ratings can be found in Table 2. Positive values indicate a preference for a CS whereas negative values indicate the opposite (the internal consistency of the positive and negative ratings was excellent; Cronbach's alpha = .98 and .97 respectively). Submitting mean ratings to a 2 (*Stimulus*: CS paired with positive vs. negative USs) × 2 (*Context Pairs*: same vs. opposite) mixed ANOVA (with the former factor within and the latter manipulated between participants) revealed a main effect for Stimulus, $F(1, 90) = 31.62$, $p < .001$, $\eta^2_{partial} = .26$, 95% CI [0.12; 0.39], $BF_{10} > 10^4$, and a two-way interaction between Stimulus and Context Pair Type, $F(1, 90) = 28.23$, $p < .001$, $\eta^2_{partial} = .24$, 95% CI [0.10; 0.38], $BF_{10} > 10^4$, with Bayes Factors strongly supporting the hypothesis that an EC

[5] Note that including the data for all participants in the analyses did not result in a shift in significance for any of the reported effects (in Experiments 1 or 2). That said, we decided to continue excluding these participants to be consistent with our initial data-analytic plan.

effect emerged and that it was moderated by the type of context pairs that participants encountered. Participants exposed to *identical context pairings* showed an assimilative EC effect: they liked CS1 more than CS2, $t(43) = 12.01$, $p < .001$, $d = 1.79$, 95% CI [1.31; 2.26], $BF_{10} > 10^4$. Those exposed to *opposite context pairings* showed no preference for CS1 over CS2, as the test of difference did not reach significance, $t(47) = 0.18$, $p = .86$, $d = 0.03$, $BF_{01} = 6.28$, with Bayes Factors providing additional evidence that EC effects attenuated rather than reversed as the result of opposite context pairings.

IAT. Submitting IAT scores to a one-way ANOVA with Context Pairs as a between-subjects factor revealed a main effect of Context Pair Type, $F(1, 91) = 18.53$, $p < .001$, $\eta^2_{partial} = .17$, 95% CI [0.05; 0.30], $BF_{10} > 10^2$, indicating that the presence of context pairs also moderated IAT effects. Participants who encountered *identity context pairings* showed an assimilative EC effect: they liked CS1 relatively more than CS2. Those exposed to *opposite context pairings* showed no preference for either stimulus. Whereas the former score was significantly different from zero, $t(43) = 6.73$, $p < .001$, $d = 1.01$, 95% CI [0.64; 1.38], $BF_{10} > 10^4$, the latter was not, $t(47) = 0.56$, $p = .58$, $d = .08$, $BF_{01} = 5.50$, with IAT scores attenuated rather than reversed following opposite context pairings.

Implicit-explicit correlations and contingency memory. The contingency memory task consisted of two questions: one probing for the contingency between CS1 and positive USs and another probing for the contingency between CS2 and negative USs. Participants who answered these two questions correctly were assigned a score of 1 whereas those who failed to do so were assigned a score of 0. We then assessed whether contingency memory performance was correlated with implicit and explicit evaluations (and whether the latter scores also correlated with one another). Self-reported ratings of CS1 were negatively correlated with those of CS2 ($r = -.89$) and positively correlated with IAT scores ($r = .59$) (CS2 evaluations were negatively correlated with IAT scores; $r = -.61$). Explicit evaluations

of CS1 (r = .29), CS2 (r = -.35) and IAT scores (r = .24) were all correlated with contingency memory performance (all ps < .03).

Discussion

Results indicate that EC effects are moderated by the presence of contextual pairings. Assimilative effects emerged when context trials contained identical stimuli (i.e., participants explicitly and implicitly preferred the CS paired with positive over the CS paired with negative images). Yet exposure to context trials containing paired events that were opposite to one another obliterated EC effects, with implicit and explicit evaluations attenuated to non-significance.

Experiment 2

Experiment 2 set out to replicate and extend our initial findings in two ways. First, we now manipulated not only the nature (same vs. opposite) but also valence of the context pairings. One group of participants encountered valenced context pairs (e.g., *happy-sad*) whereas another group were exposed to non-valenced pairs (e.g., *right-left*). On the one hand, valenced context pairs might be seen as more similar to the CS-US pairs (which also involved one valenced stimulus) that might increase the probability that the nature of the relation on context trials was seen as diagnostic for the symbolic meaning of the pairings on the CS-US trials. On the other hand, non-valenced context pairs might convey more clearly the oppositional nature of the context pairs, thereby increasing the power of those pairs as a way of changing the symbolic meaning of the CS-US pairings also. Therefore, although there are reasons to expect that this manipulation would influence the magnitude of the context effects on EC, we did not have clear *a priori* predictions regarding the direction of its effect. Second, we presented context trials and CS-US trials within the same block rather than in different blocks, hoping that this would highlight that both types of trials occur in the same temporal and spatial context and thus increase the likelihood that people generalize the relational

meaning of the context trials to the CS-US trials. We hoped that by implementing these two procedural changes we could strengthen the impact of our manipulation on resulting EC effects, and thus move from an attenuation to a full reversal in evaluations.

Method

Participants and Design. Two hundred and fourteen participants (132 women) ranging from 18 to 53 years (M = 33.9, SD = 8.9) participated online via the Prolific Academic website in exchange for a monetary reward (€1.25). The experiment involved a two factor between-subjects design: *Context Pair Type* (same vs. opposite) and *Context Pair Valence* (valenced vs. non-valenced), with self-reported ratings and IAT scores as the main dependent variables. Evaluative task order and IAT block order were also manipulated between participants. The sample size was determined prior to data collection on a convenience sampling manner.

Materials. The same CSs were used as in Experiment 1. Eight positive (*friend, happy, healthy, clean, selfless, freedom, pleasure, nice*) and eight negative adjectives (*enemy, sad, sick, dirty, selfish, imprisoned, pain, nasty*) served as USs. The same USs were also used during context pairings. The following words served as non-valenced stimuli during context pair trials: *big, small, loud, quiet, black, white, fire, ice, heavy, light, old, young, fast, slow, up, down, hot, cold, left, right, day, night*.

Procedure

There were three experimental phases: a learning phase, evaluative measures, and exploratory questions.

Learning phase. Participants were first informed that they would see two words on the screen and that they should pay attention to what appeared at all times. The learning phase consisted of four blocks that each contained 10 mini-blocks (40 mini-blocks in total). Each mini-block was comprised of two separate trials: the presentation of a context pair

followed by the presentation of a CS-US pair. A rectangular frame was presented onscreen at trial-onset. After 1000ms the first word of the context pair appeared, and 1000ms thereafter the second word of the context pair was presented. Both remained onscreen for a further 2000ms before disappearing. After a 1000ms intra-trial interval a CS appeared and remained alone onscreen for 1000ms. Thereafter a US appeared and the two stimuli remained onscreen for a further 2000ms. All stimuli then disappeared, and following a 2000ms inter-trial-interval, the next mini-block began.

On EC trials, CS1 was presented with one of eight different positively valenced adjectives whereas CS2 was presented with one of eight negative adjectives. On context trials, participants in the *identical valenced context pairs* condition encountered identically valenced words (e.g., *Happy-Happy, Sad-Sad*) whereas their counterparts in the *non-valenced identical context pairs* condition encountered identical non-valenced words (e.g., *Small-Small, Left-Left*). Those in the *opposite valenced context pairs* condition were exposed to valenced words that were opposite in meaning (e.g., *Happy-Sad, Love-Hate*) and the *non-valenced opposite context pair* condition encountered non-valenced words with opposite meanings (e.g., *Right-Left, Fast-Slow*).

Evaluative measures. Self-reported ratings and automatic evaluative responding (IAT) were assessed as in Experiment 1.

Exploratory questions. A similar set of exploratory questions were administered as in Experiment 1. We also asked participants to complete a relational matching-to-sample procedure and a need for cognition scale (Cacioppo & Petty, 1984). Again these variables were registered purely for exploratory purposes and will not be discussed further.

Results

Data Preparation

Participants who failed to complete the whole experiment were excluded (N = 12). Of those who did complete, we omitted a further nine who had IAT error rates above 30% across the entire task (n = 4), above 40% for any one of the four critical blocks (n = 2), or who responded faster than 400ms on more than 10% of trials (n = 3). This led to a final sample of 193 participants.

Preliminary Analyses

Descriptive analyses for hypothesis and context pair awareness/influence as well as contiguity memory can be found in Table 1.

Hypothesis Testing

Self-reported ratings. Mean ratings can be found in Table 3. Submitting evaluative scores to a 2(*Stimulus*) x 2 (*Context Pair Valence*) × 2 (*Context Pair Type*) mixed ANOVA (with the first factor manipulated within and the latter two factors between participants) revealed a main effect for Stimulus, $F(1, 189) = 20.83$, $p < .001$, $\eta^2_{partial} = .10$, 95% CI [0.03; 0.18], $BF_{10} > 10^4$, with Bayes Factors offering strong support for the hypothesis that an EC effect emerged. We also observed a two-way interaction between Stimulus and Context Pair Type, $F(1, 189) = 17.14$, $p < .001$, $\eta^2_{partial} = .09$, 95% CI [0.02; 0.16], $BF_{10} > 10$, as well as a three-way interaction between Stimulus, Context Pair Valence, and Context Pair Type, $F(1, 189) = 12.11$, $p = .001$, $\eta^2_{partial} = .06$, 95% CI [0.01; 0.14], $BF_{10} > 10$. To specify this three-way interaction we consider the impact of context pair type separately for those in the valence and non-valenced context pair conditions.

Participants exposed to valenced contextual pairs showed a main effect of Stimulus, $F(1, 85) = 7.46$, $p < .01$, $\eta^2_{partial} = .08$, 95% CI [0.01; 0.20], $BF_{10} > 8.3$, as well as a two-way interaction between Stimulus and Context Pair Type, $F(1, 85) = 22.91$, $p < .001$, $\eta^2_{partial} = .21$, 95% CI [0.08; 0.35], $BF_{10} > 10^3$. Paired-sample t-tests indicated that an assimilative EC effect emerged in the *identical context pairings* condition: participants liked CS1 more than CS2,

$t(41) = 6.17$, $p < .001$, $d = 0.95$, 95% CI [0.58; 1.31], $BF_{10} > 10^4$. In absolute terms, a contrastive EC effect emerged in the *opposite context pairings* condition: participants liked CS2 more than CS1. Note, however, that the difference between CS2 and CS1 ratings did not reach significance in this latter condition, $t(44) = 1.32$, $p = .20$, $d = 0.19$, 95% CI [-0.10; 0.49], $BF_{01} = 2.77$, thus indicating that EC effects were attenuated rather than completely reversed following opposite context pairs. Although participants exposed to non-valenced context pairings also showed a main effect of Stimulus, $F(1, 104) = 14.49$, $p < .001$, $\eta^2_{partial} = .12$, 95% CI [0.03; 0.24], $BF_{10} > 10^3$, they did not show any main or interaction effects for Context Pair Type (*ps* > .32), such that CS1 was always liked more than CS2 regardless of the type of context pairs encountered.

IAT. Submitting IAT scores to a 2 (*Context Pair Type*) × 2 (*Context Pair Valence*) ANOVA revealed a main effect for Context Pair Valence, $F(1, 191) = 6.06$, $p = .02$, $\eta^2_{partial} = .03$, 95% CI [0.01; 0.09], $BF_{10} = 2.37$, and Context Pair Type, $F(1, 191) = 14.66$, $p < .001$, $\eta^2_{partial} = .07$, 95% CI [0.02; 0.15], $BF_{10} > 10$, but no interaction between the two, $F(1, 191) = 3.99$, $p = .07$, $\eta^2_{partial} = .02$, 95% CI [0.00; 0.08], $BF_{10} = 1.08$. Participants showed an IAT effect favoring CS1 over CS2 in the non-valenced condition, indicated by an IAT effect that was significantly different from zero, $t(104) = 2.35$, $p = .02$, $d = 0.23$, 95% CI [0.04; 0.45], $BF_{10} = 1.5$, and no such effect in the valenced context pair condition, $t(86) = -1.21$, $p = .23$, $d = -0.13$, 95% CI [-0.34; 0.08], $BF_{01} > 4.18$. Perhaps more importantly, for the current paper, they showed an assimilative EC effect in the *identical context pairings* condition (i.e., they liked CS1 more than CS2), $t(91) = 3.17$, $p = .002$, $d = 0.33$, 95% CI [0.12; 0.54], $BF_{10} = 11.80$, and a tendency for a contrast effect in the opposite context pairings condition (i.e., they tended to liked CS2 more than CS1). Note that this latter effect failed to significantly differ from zero, $t(99) = 1.93$, $p = .06$, $d = 0.19$, 95% CI [-0.06; 0.39], $BF_{01} = 1.52$, with IAT scores attenuated rather than completely reversed following opposite context pairs.

Implicit-explicit correlations and contingency memory. Self-reported ratings of CS1 were negatively correlated with those of CS2 (r = -.49) and positively correlated with IAT scores (r = .44) (CS2 evaluations were negatively correlated with IAT scores; r = -.51). Explicit evaluations of CS1 (r = .35), CS2 (r = -.39) and IAT scores (r = .39) were all correlated with contingency memory performance (all ps < .001).

Discussion

We once again found that context pairings moderated EC effects. Similar to Experiment 1, assimilative EC effects emerged on explicit and implicit measures whenever people encountered context trials containing identical stimuli. Yet those same effects were once again attenuated when context trials contained stimuli that were opposite to one another. Interestingly these effects were evident when the context pairs were valenced in nature but largely absent when the context pairs were non-valenced.

General Discussion

We recently introduced a new symbolic perspective on EC that consists of three ideas: (a) pairings represent a contextual cue in the environment, (b) humans treat this cue as a symbol indicating that the CS and US are related in a certain way, and (c) it is this symbolic relationship between stimuli – established by pairings – which determines the subsequent change in liking. Whereas past work could be seen as relying on direct and explicit manipulations to alter the meaning of pairings, such as instructions (Moran & Bar-Anan, 2015; Peters & Gawronski, 2011; Zanon et al., 2014) and relational qualifiers (Förderer & Unkelbach, 2012), we adopted a different approach. One hypothesis to fall out of our symbolic perspective is that contextual pairings should moderate EC effects.

Across two studies we asked participants to complete a learning phase comprised of two types of trials: context trials in which two (valenced or non-valenced) words were identical or opposite to one another, and EC trials, where a CS was paired with a US. In both

cases we found that exposure to identical context pairs led to assimilative EC effects whereas opposition-based pairs led to reduced EC effects on explicit and implicit measures. These results are consistent with the prediction that EC is a function of the relational implications of stimulus pairs presented in the same context as CS-US pairs.

Implications for a symbolic perspective on EC

Strong vs. weak symbolic account. The fact that context pairs (just like instructions and relational qualifiers) can be used to moderate EC effects raises an interesting question: does the fact that pairings *can* serve as a relational contextual cue mean that EC effects are (by default) driven by pairings acting in such a manner? At this early stage one could respond to this question by referring to the two different versions of our symbolic account that we discussed in the introduction – a weak and a strong version. The *weak symbolic perspective* states that - under certain conditions - pairings can either function as a symbol or as a mere *proximal cause* of the change in liking. The goal then of EC research (given that we and others have found that pairings can function as a relational cue) is to identify and explain how, when, and why pairings function as a mere proximal cause vs. symbolic cue for liking (see De Houwer & Hughes, 2016 for more as well as concrete recommendations on how to do so).

In contrast, a *strong symbolic perspective* would argue that pairings always function as a symbol. Although it is relatively easier to test whether pairings are functioning as a symbol (i.e., to provide confirmatory evidence in support of a strong account), it seems relatively more difficult to show that pairings are not functioning as a symbol in a given context (i.e., to disprove the strong account) (for reasons why see De Houwer & Hughes, 2016). As such, we foresee a difficult and lengthy debate about whether EC can be found in the total absence of symbolic meaning construction.

Attenuation vs. reversal. As we outlined in the introduction, the addition of relational qualifiers or instructions implying an opposition relation between paired stimuli often leads to attenuated rather than reversed EC effects (e.g., Fiedler & Unkelbach, 2010; Hu et al., 2017; Moran, Bar-Anan, & Nosek, 2016; Peters & Gawronski, 2011). We observed a similar pattern when context-pairs were used. On first glance, an attenuation rather than reversal in liking may seem to support a dual-process perspective while refuting a symbolic one. From a dual-process perspective, pairings may lead to the formation of an association between the CS and US (resulting in an assimilative effect) while context pairs lead to a proposition about *how* those stimuli are related (resulting in a contrastive effect), with the former process still exerting an impact on evaluations despite the presence of the latter. Yet an attenuation effect could also be accommodated by the aforementioned symbolic accounts in two ways.

The first is a within-subjects account. In this case attenuation could be due to the fact that a small amount of training with a limited number of context pairs does not completely override the impact of a long learning history in which pairings function as a relational cue for similarity rather than opposition. That is, it may take more than a short 3-5 minute intervention to fully transform the 'default' symbolic meaning of pairings, especially with a 'subtle' manipulation as used here. This might also explain why others have failed to find a full reversal (when using relational qualifiers and/or instructions) given that these manipulations are also exceptionally brief in nature. In other words, just as it can take time and effort to change the original symbolic meaning of a stimulus when it is overlearned (e.g., it might take some time for an English speaker to automatically respond to 'pain' as meaning 'bread' after arriving in France), so too might it take more than several minutes with a subtle manipulation to change the symbolic meaning of a regularity – especially if the original meaning of that regularity has also been overlearned in the past. Put simply, participants might certainly learn that pairings have a particular meaning in the experimental context

(opposition) and yet still recall the distal symbolic meaning of pairings that applies in many other contexts (similarity) (i.e., pairings may have two meanings for the same individual that simultaneously have an impact on liking). If so, then repeated training of the novel meaning of pairings across both time and context may be required to fully reverse evaluations.

The second explanation is a between-subjects account. This would argue that context pairings do fully override the default meaning of pairings for some participants but leave the original meaning intact for others. The result is that evaluative effects appear to absent at the overall group level. This account resonates with the current data: if we only probe for an impact of context-pairs at the overall (group) level (as was typically the case in previous work) then we do indeed observe an attenuation effect. Yet a closer inspection of the individual-level data reveals descriptively different patterns of evaluation across participants (see supplementary materials). Whereas participants in the same context pairings condition generally show effects in the expected direction, there is considerably more variation in the opposite context-pairs condition. Specifically, some participants show a zero score (on the self-reports), others produce positive values, and still others show negative values.

One possibility is that there are distinct subgroups of participants within the opposition condition. Some might be genuinely ambivalent towards the CSs (either because no evaluation was formed in the first place or because the two symbolic meanings of pairings cancel each-other out). Others might be responding in-line with the original meaning of pairings and thus show an assimilative effect (either because they failed to learn the new meaning conveyed by the context-pairs or, despite understanding this new contextual meaning, automatically make similarity-based inferences about the CS valence). Still others respond in-line with the novel meaning of pairings and thus show a contrastive effect.

The key point here is that simply reporting findings at the overall group level may serve to hide potential subgroups of participants who fully reverse, attenuate, or continue to

show the original effect. Critically, this may also be the case in previous studies which report attenuation effects and is clearly worthy of additional attention. It also highlights the need for caution when making strong claims about mental mechanisms (i.e., between dual vs. single process models of evaluation) on the basis of such findings. Inferences made on the basis of group-level data may support certain theoretical claims (dual-process accounts) whereas those made based on individual-level data support others (single-process accounts).

Alternative (non-symbolic) explanations. The attenuation effects observed in our studies could also be driven by two other factors. The context pairing task used in the current experiments attempted to change the meaning of pairings *via* stimulus pairings. Unlike instructions, such a pairing-based manipulation may reinforce the 'default' meaning of pairings (similarity) in the act of trying to alter that meaning (opposition). This is somewhat analogous to pressing the accelerator in a car while pulling the brake: participants are asked to think of pairings as a cue for opposition while the original meaning (similarity) is being repeatedly elicited each time that they encounter one stimulus being paired with another. We also tried to change the meaning of pairings during the EC phase itself. This requires participants to not only discover the new meaning of pairings but simultaneously use it to inform their stimulus evaluations. Although it worked, we may have had more success if our symbolic meaning manipulation was distinct from, and actually came before, the EC phase (e.g., similar to the priming task used by Corneille et al., 2009, or a pre-training phase that involves exclusively opposite pairs presented across several experimental sessions and only then the EC phase). Previous work indicates that timing matters when it comes to changing the meaning of pairings: providing explicit information about the meaning or validity of paired events *before* people encounter those pairings influences explicit and implicit evaluations whereas doing so *after* the pairings influences explicit but not implicit

evaluations (e.g., Gregg, Seibt, & Banaji, 2006; Peters & Gawronski, 2011; Zanon et al., 2014; although see Moran, Bar-Anan, & Nosek, 2017).

Broader implications. Conceptualizing pairings as a symbolic cue also has implications for other research areas, including classical conditioning and persuasion. With respect to conditioning, our findings support the idea that human behavior is often based on the meaning attributed to, rather than the mere physical properties of, the environment. In other words, our symbolic account is constructivist in nature and highlights that a specific type of meaning construction (symbolic) may take place via a specific type of proximal event (pairings). This may also be the case in other, non-evaluative domains, such as causal (e.g., Waldmann, 2017) and fear learning (Craske, Hermans, & Vansteenwegen, 2006). Likewise, the suggestion that only humans can use symbols (e.g., Deacon, 1997; Hayes et al., 2001) opens up yet another interesting possibility: it may be that the types of symbolic effects reported here are restricted to humans who have developed the ability to use symbols and are absent in humans and non-humans who have not.

The symbolic account also suggests that EC and persuasion may have more in common than previously thought. For instance, theories of persuasion such as the Heuristic-Systematic Model (Chaiken, Liberman, & Eagly, 1989) and Elaboration-Likelihood Model (e.g., Petty & Cacioppo, 1986) view the mental processes that mediate changes in liking due to pairings (EC) as primitive in nature. The former (HSM) would consider pairings as a heuristic cue (i.e., an environmental cue which elicits an information processing strategy based on simple rules, schemas, or prior knowledge) (e.g., pairings automatically elicit a rule such as “stimuli which co-occur are similar in valence”). The latter (ELM) typically relegates pairings to the peripheral route of attitude change, and sees it as a potential input for, but rarely a type of, argument in itself (Petty & Brinol, 2014). Other theories, like the Unimodel of persuasion, do allow for “persuasive evidence to be presented in an unlimited number of

forms and variations", (Bohner, Erb, & Siebler, 2008; p.172) including forms that do not involve words or sentences, such as pairings (Kruglanski & Thompson, 1999). In short, certain theories of persuasion tend to view pairings as a simple heuristic cue whereas others allow for the possibility that pairings could function as a persuasive argument.

Yet many types of EC effects – including those reported here - involve more than the mere pairing of stimuli. Participants are often given extra verbal information about the pairings (e.g., relational qualifiers, instructions, cover stories) and/or provided with the motivation and opportunity to think about the pairings in some way (e.g., make online judgements). This combination of verbal information and pairings (and/or requirement to effortfully think about them) may lead people to treat pairings, not as a heuristic cue, but as a 'regularity-based argument' (e.g., "the CS causes Cancer" is an argument for disliking the CS). The same may be true when (a) contextual pairings are provided along with CS-US pairings, (b) participants are given the conditions necessary to elaborate on both pieces of information, and (c) we raise their motivation to process that information via pre-task instructions (as in the current experiments). If so, then EC and persuasion may both involve changes in liking due to arguments but differ in the way that those arguments are delivered and constructed (e.g., either via stimulus pairings [EC] or words and sentences [persuasion]). Indeed, it may be that pairings - even in the absence of extra verbal instructions or information – can also serve as an argument indicating that a CS is similar to a US, and as a result, leads to a change in liking. This is not to say that pairings will always function in this regard: when motivation and ability to process pairings are low, it is likely that the mere fact that stimuli are paired does indeed function as a heuristic cue. However, increasing people's motivation and opportunities to effortfully operate on the pairings, may cause them to treat pairings as a simple, regularity-based, argument and this argument may drive their evaluations (i.e., pairings may function as both heuristic cue and persuasive argument).

Regardless, the take home message here is that EC is closely related to persuasion and compatible with several theories in the area. As far as we know, EC and persuasion have never been linked in this way, perhaps because the former is typically considered as a primitive, non-symbolic phenomenon (for more see De Houwer & Hughes, 2016; Hughes, Ye, Van Dessel, & De Houwer, in press).

Future directions. Our findings open up a number of avenues for future research on EC. First, and as we previously mentioned, researchers often view contextual information such as instructions, qualifiers, and judgments as symbolic messages that are applied *on top of* pairings. Yet our symbolic account takes a different stance - such information *alters the meaning of the pairings themselves*. This perspective makes several new predictions about EC effects. Foremost amongst these is that the symbolic meaning of pairings is not confined to same and opposite. People can – in principle – respond to pairings as a cue indicating that a CS caused or prevented, comes before or after, or is stronger or weaker than a US. Future research could modify our procedure to test this idea. For instance, we could present context pairings in which the first stimulus is a known cause of the second element (e.g., weapon – injury) or a known preventer of the second element (e.g., medicine – disease).

Second, although most participants reported that they were aware of the presence and content of the context pairs (i.e., that they involved identical or opposite stimuli), many failed to take those pairings into account when generating stimulus evaluations. Future work will need to consider whether and why certain types of context pairs, or ways of manipulating the meaning of pairings, are more effective at transferring relational properties than others. For instance, we found that non-valanced context pairs moderated EC effects in Experiment 1 and yet valenced (but not non-valenced) pairs did so in Experiment 2 (although note that the context and CS-US pairs were blocked in the former and interspersed with one another in the latter experiments). We also observed more zero scores in Experiment 2 relative to

Experiment 1. Clearly there are boundary conditions to the effectiveness of such manipulations. Future work could also directly assess what meaning the CS-US pairings have for participants rather than solely infer it from changes in evaluation. Doing so could serve as an additional manipulation check and could be correlated with specific changes in EC effects (e.g., perhaps only those who can report that the meaning of pairings has changed will show full reversals in liking). Third, future work could incorporate a control condition where no context pairings are encountered in order to determine if similarity-based context pairs increase EC effects or if opposition-based pairs significantly weaken EC effects compared to normal. Fourth, the presence or absence of context pairs at the time of encoding, storage, or retrieval may determine how much of an impact they have on the proximal meaning of CS-US pairings. Future work could test whether a similar pattern of results emerges when context pairings are provided either before or after an EC phase.

Finally, it is worth considering alternative explanations for our findings. For instance, it may be that the observed assimilative and attenuated effects were not due to the symbolic meaning of pairings established by the context trials but rather due to pairings taking place *between* (rather than *within*) trials. To illustrate, imagine that participants first encounter one stimulus pair (*Love-Hate*) followed by a second (*Love-CS1*). It may be that people ignored the contiguity between Love and CS1 and instead decided to relate CS1 directly with Hate. This alternative account does not require that people consider the symbolic implications of pairings (e.g., that paired stimuli are opposite; therefore CS1 is opposite to love). Instead, they simply need to recognize the relation between the elements of different pairs (e.g., CS1 has the same meaning as hate because they both co-occur with love) rather than the relation between the elements within a pair (*Love-Hate; CS1-Love*). Although a possible explanation for Experiment 2, it is difficult to see how it could adequately account for Experiment 1, where context pairs were non-valenced and blocked rather than interspersed with CS-US

trials. A second possibility is that pairing a positive with a negative US (e.g., Love-Hate) led to US-revaluation, where the respective USs decreased in their valence. When subsequently paired with the CS this would explain why there was an attenuated rather than reversed EC effect. Yet, once again, this hypothesis can only account for the findings obtained from our second and not first study. Future work could test both accounts by having participants (a) rate the valence of the USs before and after the learning phase and seeing if there is a change, or (b) ensure that there was no stimulus overlap between context and EC trials.

Conclusion

Our results lend further support to the idea that EC effects can be moderated by contextual (relational) information that is relevant to the paired events. Whereas past work relied on direct manipulations that added a separate message on top of pairings, we show that contextual pairings can also change the impact of CS-US pairings on implicit and explicit evaluations.

References

Briñol, P., Petty, R. E., & McCaslin, M. J. (2009). Changing attitudes on implicit versus explicit measures. In R. E. Petty, R. H. Fazio, P. Briñol (Eds.), *Attitudes. Insights from the new implicit measures* (pp. 285–326). New York: Psychology Press.

Corneille, O., Yzerbyt, V. Y., Pleyers, G., & Mussweiler, T. (2009). Beyond awareness and resources: Evaluative conditioning may be sensitive to processing goals. *Journal of Experimental Social Psychology*, *45*(1), 279-282.

Craske, M. G., Hermans, D., & Vansteenwegen, D. (Eds.). (2006). *Fear and learning: From basic processes to clinical implications*. Washington, DC: American Psychological Association.

Deacon, T. W. (1997). *The symbolic species*. New York; Norton.

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*(1), 1-20.

De Houwer, J. (2014). A propositional model of implicit evaluation. *Social and Personality Psychology Compass*, *8*(7), 342-353.

De Houwer, J., Barnes-Holmes, D., & Moors, A. (2013). What is learning? On the nature and merits of a functional definition of learning. *Psychonomic Bulletin & Review*, *20*(4), 631-642.

De Houwer, J., & Hughes, S. (2016). Evaluative conditioning as a symbolic phenomenon: On the relation between evaluative conditioning, evaluative conditioning via instructions, and persuasion. *Social Cognition*, *34*(5), 480-494.

De Houwer, J., Teige-Mocigemba, S., Spruyt, A., & Moors, A. (2009). Implicit measures: A normative analysis and review. *Psychological Bulletin, 135*(3), 347-368.

Fiedler, K., & Unkelbach, C. (2011). Evaluative conditioning depends on higher order encoding processes. *Cognition & Emotion*, *25*(4), 639-656.

Förderer, S., & Unkelbach, C. (2012). Hating the cute kitten or loving the aggressive pit-bull: EC effects depend on CS–US relations. *Cognition & Emotion*, *26*(3), 534-540.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology*, *90*(1), 1-20.

Hu, X., Gawronski, B., & Balas, R. (2017). Propositional versus dual-process accounts of evaluative conditioning: I. The effects of co-occurrence and relational information on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *43*(1), 17-32.

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational Frame Theory: A Post-skinnerian account of human language and cognition*. New York: Plenum Press.

Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: a meta-analysis. *Psychological Bulletin, 136*(3), 390-421.

Hughes, S., De Houwer, J., & Barnes-Holmes, D. (2016). The Moderating Impact of Distal Regularities on the Effect of Stimulus Pairings. *Experimental Psychology*, 63, 20-44.

Hughes, S. J., Ye, Y., Van Dessel, P., & De Houwer, J. (*in press*). When People Co-occur with Good or Bad Events: Graded Effects of Relational Qualifiers on Evaluative Conditioning. *Personality & Social Psychology Bulletin*.

Moran, T., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743-752.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2015). Processing goals moderate the effect of co -occurrence on automatic evaluation. *Journal of Experimental Social Psychology, 60*, 157-162.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2016a). The assimilative effect of co-occurrence on evaluation above and beyond the effect of relational qualifiers. *Social Cognition*, *34*(5), 435-461.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The Effect of the Validity of Co -occurrence on Automatic and Deliberate Evaluation. *European Journal of Social Psychology*, 47(6), 708-723.

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557-569.

Schwarz N. 1994. Judgment in a social context: Biases, shortcomings and the logic of conversation. In *Advances in Experimental Social Psychology* (Vol. 26), Zanna, M. (ed.). Academic Press: San Diego, CA; 123–162.

Unkelbach, C., & Fiedler, K. (2016). Contrastive CS-US relations reverse evaluative conditioning effects. *Social Cognition*, *34*(5), 413-434.

Waldmann, M. R. (Ed.) (2017). *The Oxford Handbook of Causal Reasoning*. New York: Oxford University Press.

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning?. *The Quarterly Journal of Experimental Psychology*, *67*(11), 2105-2122.

Appendix

Table 1. Descriptive statistics for hypothesis awareness and influence, context pair awareness and influence, and contiguity memory in Experiments 1 and 2.

	Experiment 1		Experiment 2	
	Yes	No	Yes	No
Hypothesis Awareness	2 (2.2%)	90 (97.8%)	10 (5%)	183 (94%)
Hypothesis Influence	8 (8.7%)	84 (91.3%)	6 (3%)	187 (96%)
Context Pair Awareness	82 (89.1%)	6 (6.5%)	193 (100%)	0 (0%)
Context Pair Influence	40 (43.5%)	49 (53.3%)	65 (33%)	128 (66%)
Contingency Memory	67 (72.8%)	22 (23.9%)	123 (64%)	70 (36%)

Note. Participants were defined as hypothesis aware if they indicated that we used the context pairings to influence how they responded and as unaware if they did not. They were defined as hypothesis influenced if they said that the context pairings influenced their CS evaluations. Participants were defined as context pair aware if they noticed the relational nature of the context pairs during the learning phase (e.g., that we presented words that were similar or opposite to one another) and influence aware if they thought the pairs influenced how they evaluated the CSs. Participants correctly remembered the contingency between CSs and USs if they stated that CS1 was paired with positive and CS2 paired with negative USs.

Table 2. Means and standard deviations of self-reported ratings and IAT scores as a function of context pair type and valence in Experiment 1.

Pair Type	Same (N = 44)	Opposite (N = 48)	Overall (N = 92)
Ratings			
CS1	3.29 (1.97)	-0.06 (3.85)	1.54 (3.51)
CS2	-3.53 (1.99)	-0.25 (3.83)	-1.82 (3.49)
IAT D4 scores	0.44 (0.44)	0.04 (0.49)	0.23 (0.51)

Table 3. Means and standard deviations of self-reported ratings and IAT scores as a function of context pair type and valence in Experiment 2.

	Valenced Pairs			Non-Valenced Pairs		
Pair Type	Same (N = 42)	Opposite (N = 45)	Overall (N = 87)	Same (N = 50)	Opposite (N = 56)	Overall (N = 106)
Ratings						
CS1	2.19 (2.19)	0.22 (3.04)	1.17 (2.83)	1.09 (1.72)	1.14 (2.52)	1.12 (2.17)
CS2	-1.43 (2.27)	1.21 (2.74)	-0.07 (2.84)	-0.55 (2.17)	-0.10 (2.71)	-0.31 (2.47)
IAT D4 scores						
	0.14 (0.55)	-0.28 (55)	-0.08 (0.59)	0.19 (0.47)	0.04 (0.49)	0.11 (0.48)

Supplementary Materials

Table 1. Number of participants who showed positive, negative, or neutral evaluative effects in Experiments 1-2 towards CS1 (paired with positive USs), CS2 (paired with negative USs), or on the IAT, as a function of context-pairing type (same or opposite).

	Experiment 1		**Experiment 2**	
	Same	*Opposite*	*Same*	*Opposite*
Self-Reported Ratings (CS1)				
Positive effect	39 (88%)	24 (50%)	59 (64%)	49 (49%)
Neutral effect	3 (7%)	3 (6%)	21 (23%)	24 (24%)
Negative effect	2 (5%)	21 (44%)	12 (13%)	28 (28%)
Self-Reported Ratings (CS2)				
Positive effect	1 (2%)	22 (46%)	21 (23%)	45 (45%)
Neutral effect	3 (7%)	3 (6%)	25 (27%)	27 (27%)
Negative effect	40 (91%)	23 (48%)	46 (50%)	29 (29%)
IAT effect				
Positive effect	36 (82%)	24 (50%)	60 (65%)	46 (46%)
Neutral effect	1 (2%)	3 (6%)	0 (0%)	0 (0%)
Negative effect	7 (16%)	21 (44%)	32 (35%)	54 (54%)

Same context-pairs condition

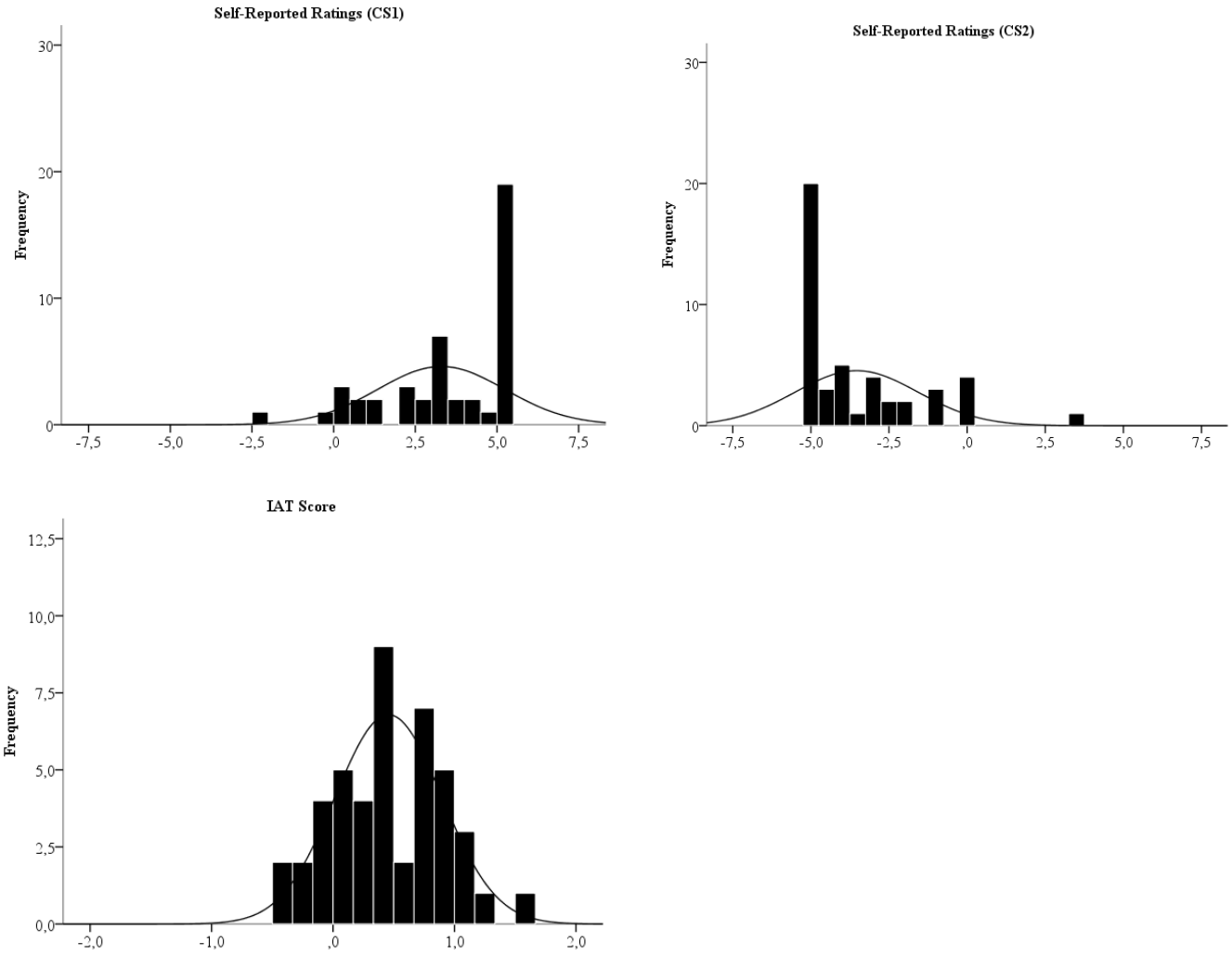

Figure 1. Frequency distribution of individual-level effects for the self-reported ratings of CS1 and CS2 (as well as IAT scores) for those in the same context pairs conditions in Experiment 1.

Opposite context-pairs condition

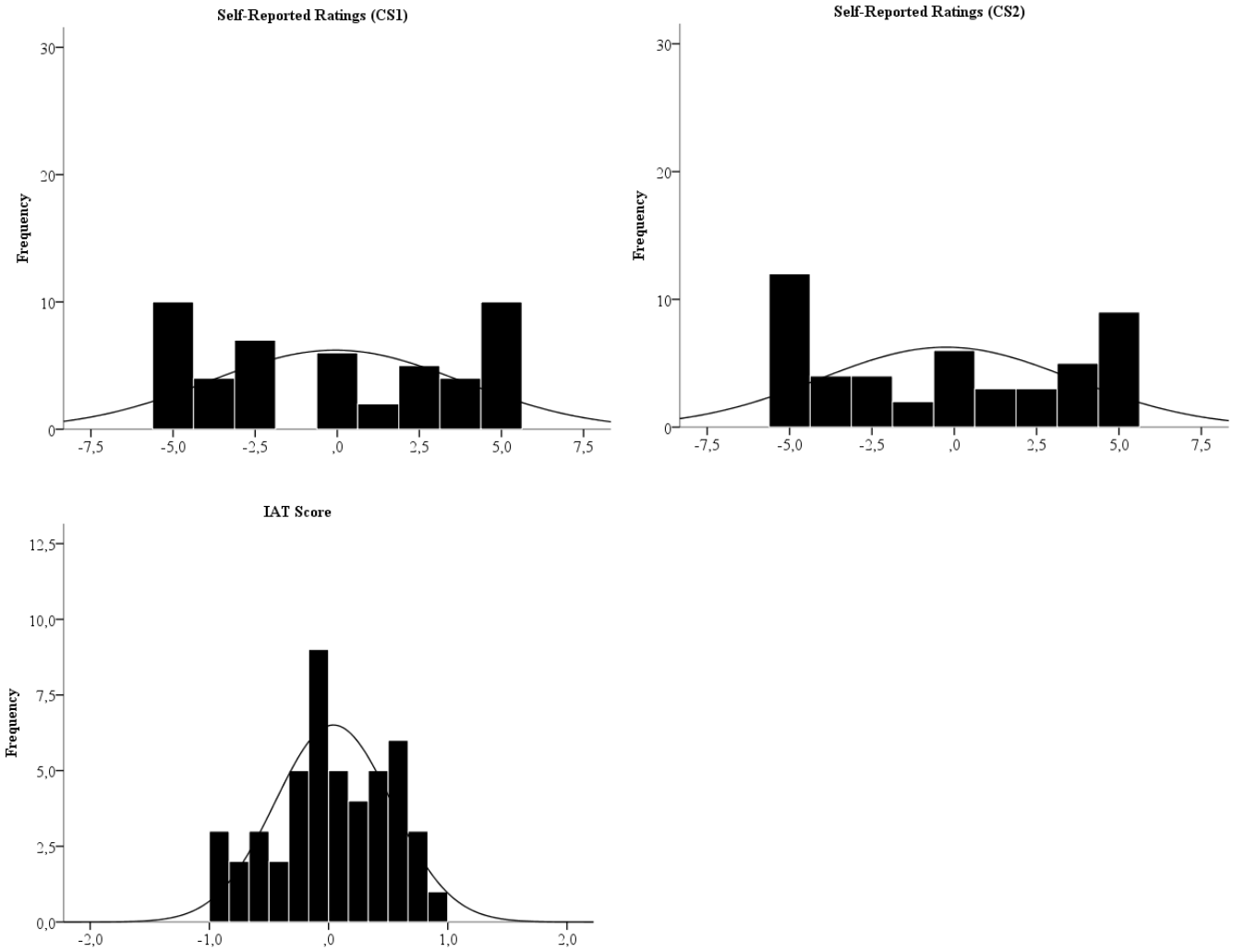

Figure 2. Frequency distribution of individual-level effects for the self-reported ratings of CS1 and CS2 (as well as IAT scores) for those in the opposite context pairs conditions in Experiment 1.

Same context-pairs condition

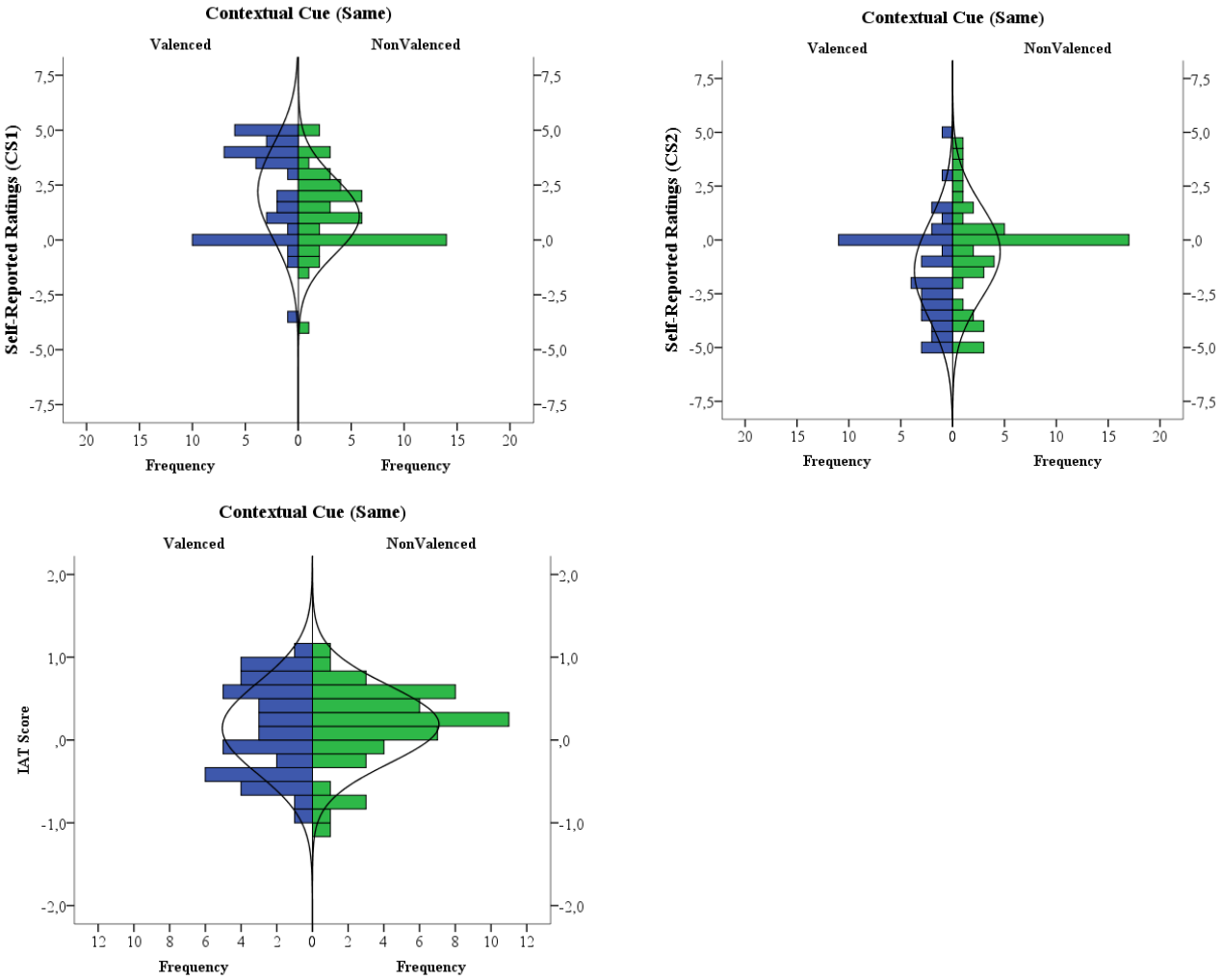

Figure 3. Frequency distribution of individual-level effects for the self-reported ratings of CS1 and CS2 (as well as IAT scores) for those in the valenced and non-valenced same context pairs conditions in Experiment 2.

Opposite context-pairs condition

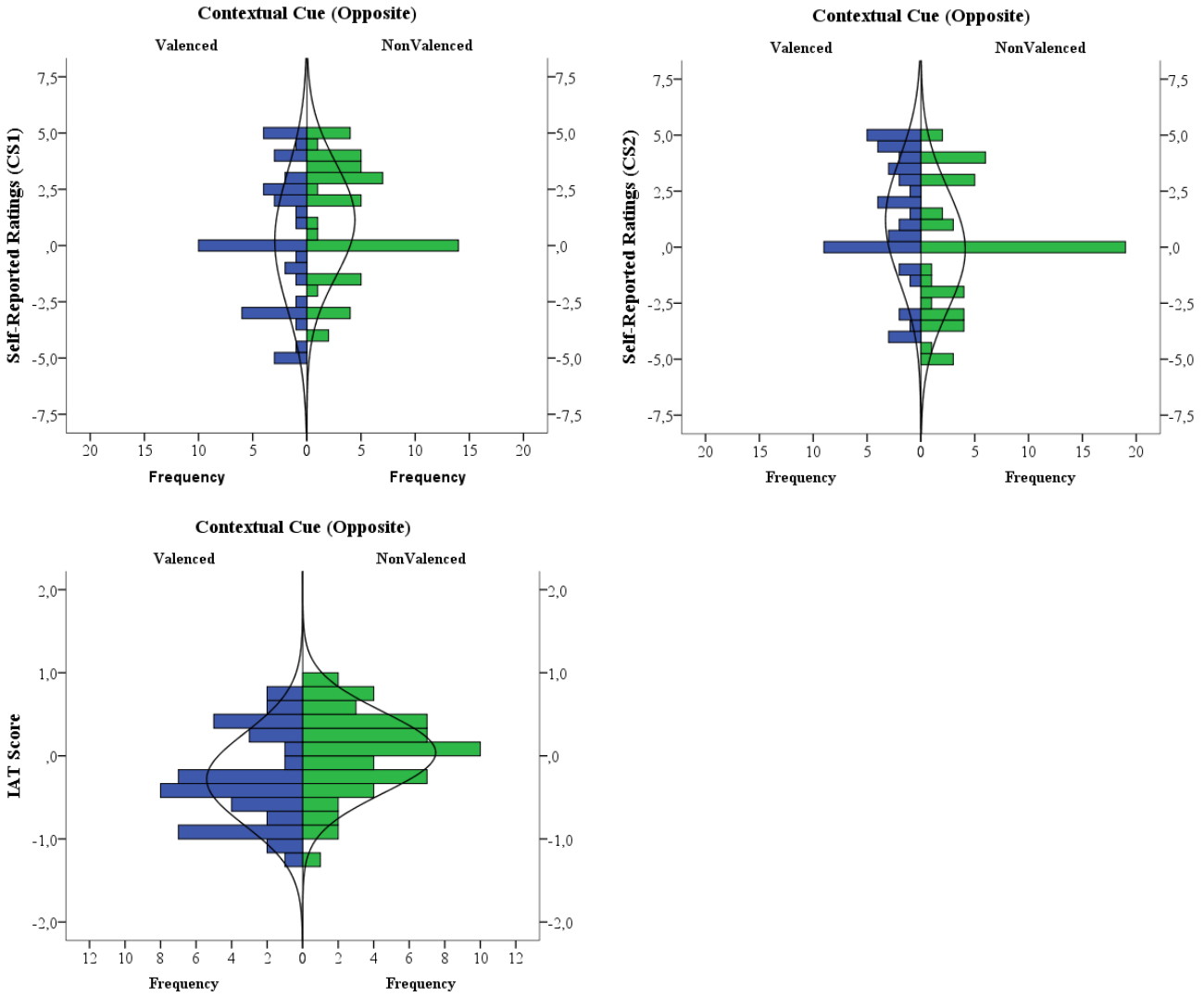

Figure 4. Frequency distribution of individual-level effects for the self-reported ratings of CS1 and CS2 (as well as IAT scores) for those in the valenced and non-valenced opposite context pairs conditions in Experiment 2.